Meta Llama 4 Launches April 5 with 109B-Parameter Scout Model—17B Active Parameters Fit on Single H100, Outperform GPT-4.5 and Claude 3.7 on STEM Benchmarks

Meta just made frontier AI fit on one GPU—and simultaneously banned 450 million Europeans from using it.

The News: Meta’s Mixture of Experts Bet Arrives

On April 5, 2025, Meta released Llama 4, marking the company’s first deployment of Mixture of Experts (MoE) architecture in their flagship open model line. The release includes three variants: Scout (available now), Maverick (available now), and Behemoth (still training).

The headline numbers tell the story. Scout carries 109 billion total parameters but

activates only 17 billion during inference—a 6.4x efficiency multiplier that lets it run on a single NVIDIA H100 GPU with Int4 quantization. Maverick scales to 400 billion total parameters across 128 experts while maintaining the same 17 billion active parameter footprint.

Both models shipped simultaneously on Llama.com and Hugging Face, with immediate integration into Meta AI across WhatsApp, Messenger, and Instagram in 40 countries. The catch: EU-domiciled users and companies are explicitly banned from using Llama 4 under the updated Community License, a regulatory standoff affecting over 450 million potential users.

Why This Matters: The Economics of Open AI Just Changed

The real story isn't the benchmark scores—it's the infrastructure economics.

Running frontier models has historically required multi-GPU clusters that pushed inference costs beyond what most organizations could justify for production workloads. A single H100 costs roughly $30,000; a cluster of eight runs $240,000 before you account for networking, cooling, and redundancy. Scout changes this calculus by delivering GPT-4-class performance on hardware most AI-focused companies already own.

This creates three immediate winners. First, mid-market enterprises that couldn't justify eight-figure infrastructure investments now have a viable path to running frontier models in-house. Second, inference providers like Together AI, Anyscale, and Fireworks can offer dramatically lower per-token pricing while maintaining margins. Third, any company running sensitive workloads—healthcare, legal, financial services—gains a real option for keeping data on-premises without sacrificing capability.

The losers are equally clear. OpenAI and Anthropic face renewed pricing pressure at the API layer. According to TechCrunch's coverage, Scout and Maverick outperform GPT-4.5, Claude Sonnet 3.7, and Gemini 2.0 Pro on MATH-500 and GPQA Diamond benchmarks—the two most respected evaluations for STEM reasoning. If these numbers hold under independent testing, the "quality premium" that justified proprietary API pricing starts to evaporate.

European AI startups face a more existential problem. The EU ban isn't a minor inconvenience—it's a competitive disability. While American and Asian competitors build on Llama 4, European teams must either find workarounds (likely violating the license), stick with Llama 3, or pay for proprietary alternatives. The irony of AI regulation designed to protect European interests actively disadvantaging European companies will not be lost on Brussels.

Technical Deep Dive: How Mixture of Experts Actually Works Here

The MoE Architecture Explained

Mixture of Experts isn't new—Google deployed it in the Switch Transformer back in 2021—but Meta's implementation in Llama 4 represents the first time MoE has appeared in a widely-distributed open model at this scale.

The core concept: instead of passing every token through every parameter (dense architecture), MoE models route each token to a subset of specialized "expert" subnetworks. Scout uses 16 experts with 17 billion parameters each; Maverick scales to 128 experts. A learned routing network decides which experts process each token, typically activating 2-4 experts per forward pass.

This is why "109 billion parameters" and "17 billion active parameters" aren't contradictory—they're describing different things. Total parameters represent the full model weight; active parameters represent what actually computes during inference. Your GPU only needs to handle the active parameters plus the routing overhead.

The tradeoff is memory bandwidth. Even though compute remains constant, all 109 billion parameters must be accessible in memory for the router to select experts. This is why Int4 quantization matters—it compresses each parameter from 16 bits (FP16) to 4 bits, cutting memory requirements by 4x. Scout's 109 billion parameters at Int4 require roughly 55GB of memory, fitting comfortably within an H100's 80GB capacity.

The 10 Million Token Context Window

Scout supports a 10 million token context length—roughly 7.5 million words, or

about 75 average-length novels. This number sounds impressive until you examine the practical constraints.

Context length and context usability are different things. Processing 10 million tokens requires loading and attending to all previous tokens for each new generation, creating quadratic computational scaling. In practice, useful context with current attention mechanisms caps out much lower due to attention dilution—the model struggles to maintain retrieval accuracy for information buried deep in the context.

Meta hasn't published detailed benchmarks on retrieval accuracy across the full 10 million token window. Until independent testing confirms consistent performance at scale, treat this number as a theoretical maximum rather than a production specification. For most workloads, effective context likely caps between 100,000 and 500,000 tokens—still substantial, but not the headline figure.

That said, extended context opens real applications: full-codebase analysis, multi-document synthesis, and long-running agentic workflows where maintaining state across sessions matters. The architecture supports it; the question is how well.

Benchmark Performance in Context

The benchmark comparisons focus on two evaluations: MATH-500 (500 competition-level math problems) and GPQA Diamond (graduate-level science questions written by PhD researchers). Meta claims Behemoth—still training—outperforms GPT-4.5, Claude Sonnet 3.7, and Gemini 2.0 Pro on both.

Note the careful wording. The benchmark claims apply to Behemoth, not to the currently-available Scout and Maverick. Meta's blog states Scout and Maverick are "competitive" with frontier models, but the superlative claims attach to a model not yet released. This is important context that some coverage has blurred.

Independent benchmark reproduction typically takes 2-4 weeks as the community runs standardized evaluation suites. Until then, treat performance claims as directionally accurate but unverified. Meta's track record with Llama 3 benchmarks was solid; expect similar accuracy here, with potential overstatement at the margins.

The Contrarian Take: What Everyone's Getting Wrong

Overhyped: "Open Source" Framing

Llama 4 is not open source. It's source-available under a proprietary license with significant restrictions.

The Llama 4 Community License bans EU users outright. It requires companies with over 700 million monthly active users to obtain special Meta approval—effectively blocking competitors like TikTok, Snap, and any scaled consumer platform from unrestricted use. Derivative models must include "Llama" in their names. Commercial use requires license acceptance.

Compare this to actually-open models like Mistral's releases under Apache 2.0 or the various community efforts under MIT licensing. Llama 4's license exists to advance Meta's strategic interests: commoditizing the AI layer to undermine competitors while maintaining enough control to prevent direct competitive threats. This isn't criticism—it's accurate framing. Call it "openly available" or "source-available," but "open source" misleads.

Underhyped: The Multimodal Pre-training

Coverage has focused on parameters and benchmarks while largely ignoring that Scout and Maverick are natively multimodal. These models process text, images, and video as first-class inputs, pre-trained together rather than bolted on after the fact.

This matters more than incremental benchmark improvements. Native multimodality means the model's internal representations unify across modalities—it doesn't translate images to text internally before reasoning. For applications like document understanding, video analysis, and robotics, this architectural choice compounds over time as fine-tuning and specialization build on a more capable foundation.

The models that win in 2026 will be the ones that best integrated modalities in 2025. Llama 4's architecture positions it well for that future even if current benchmarks don't fully capture the advantage.

Underhyped: The 200-Language Pre-training

Meta claims 10x more multilingual tokens in Llama 4's pre-training compared to Llama 3, covering 200 languages. If true, this represents the most linguistically diverse frontier model ever released.

Most AI coverage focuses on English performance because most AI coverage is written in English. But the commercial opportunity in non-English markets is enormous and underserved. A model that performs well in Hindi, Indonesian, Arabic, Portuguese, and Swahili opens markets representing billions of users—markets where OpenAI and Anthropic have limited presence and local competitors lack frontier capabilities.

For companies building products in emerging markets, this may be Llama 4's most significant feature.

Practical Implications: What You Should Actually Do

If You're Running Inference at Scale

Start testing Scout immediately. The single-GPU footprint fundamentally changes your cost structure if performance holds.

Practical next steps: Download the Int4 quantized weights from Hugging Face. Benchmark against your current production model on your actual workloads—not standardized benchmarks, but the specific tasks your users care about. Measure latency at your p95 traffic levels. Calculate the cost delta between your current setup and Scout on single H100s.

If Scout matches your quality requirements, you're looking at potential 4-8x cost reductions depending on your current architecture. That's not incremental—it's the difference between profitable and unprofitable for many AI-native products.

If You're Building AI Products

The 700 million MAU threshold matters more than it appears. Meta structured this to allow startups and growth-stage companies unrestricted access while

maintaining leverage over scaled competitors. If your product has any chance of reaching that scale, build abstraction layers now that let you swap models without architectural rewrites.

Multimodal capabilities deserve immediate prototyping. If you've delayed features that require image or video understanding because GPT-4V or Claude's vision pricing didn't pencil out, run the numbers again with locally-hosted Llama 4. The economics may have flipped.

If You're Based in the EU

The ban creates uncomfortable choices. The license explicitly prohibits EU-domiciled entities from using Llama 4, and Meta has legal resources to enforce this.

Your options: Use Llama 3.3 (still available without geographic restrictions), evaluate Mistral's offerings (EU-based, Apache-licensed), pay for proprietary APIs from OpenAI/Anthropic, or wait for European regulatory clarity that may never come.

Do not attempt to circumvent the geographic restriction through VPNs or non-EU subsidiaries. The license is clear, Meta has enforcement incentives, and the reputational and legal risk outweighs the capability gain. This is painful advice, but it's correct advice.

If You're Evaluating Build vs. Buy

Llama 4 strengthens the "build" side of build-vs-buy for inference but doesn't change the training calculus. You still can't train frontier models without frontier capital—Meta reportedly spent billions on Llama 4's pre-training. What you can now do is run frontier-class inference on infrastructure you control at costs approaching commodity compute.

The sweet spot: use proprietary APIs for low-volume, high-variability workloads where you're still iterating on prompts and use cases. Shift to self-hosted Llama 4 for high-volume, stable workloads where you've validated quality and need to optimize costs.

The Regulatory Dimension: EU Ban Implications

Meta's decision to ban EU users isn't arbitrary—it reflects a calculated assessment that compliance costs exceed market value.

The EU AI Act imposes transparency, documentation, and audit requirements on "general purpose AI models" that Meta apparently judged incompatible with open distribution. Additionally, GDPR's restrictions on training data likely conflict with Meta's pre-training corpus construction. Rather than maintain two separate model lineages with different training data and compliance documentation, Meta chose geographic restriction.

This sets a precedent. Other model developers—particularly those releasing open weights—will face the same calculation. If compliance requires maintaining separate EU-specific models, many will follow Meta's path. The EU risks creating a two-tier AI market where European companies perpetually access capabilities 6-18 months behind their American and Asian competitors.

The counterargument: regulatory constraint forces innovation in privacy-preserving and transparent AI development, potentially yielding long-term advantages. This argument has theoretical merit but requires regulatory frameworks that actually enable compliant development of competitive models. So far, evidence for this remains thin.

Forward Look: Where This Leads

6-Month Horizon

Behemoth's release will dominate the second half of 2025. Meta's largest model, still training at announcement, represents their genuine frontier play. If Behemoth matches the claimed benchmark performance and maintains the MoE efficiency pattern, it becomes the default choice for organizations that previously defaulted to GPT-4 or Claude.

Expect aggressive pricing responses from OpenAI and Anthropic within 90 days of Behemoth's release. The proprietary moat has always been capability, not ecosystems—when capability parity arrives via open weights, pricing becomes the competitive lever.

LlamaCon on April 29, 2025 will likely reveal technical details Meta held back from the initial launch: specific architecture choices, training methodology, and roadmap items. If you're making infrastructure decisions, wait for this information before committing.

12-Month Horizon

The MoE architecture will become standard for open models. The efficiency gains are too significant to ignore, and Meta has now proven the approach works at frontier scale. Mistral, Cohere, and community efforts will follow.

More interesting: the 10 million token context window, even if not fully usable today, signals where architectures are heading. Expect dramatic progress on attention mechanisms and retrieval accuracy across extended context. The model that first achieves high-fidelity retrieval across 1 million+ tokens with acceptable latency wins the enterprise market.

Regulatory fragmentation will worsen before it improves. The EU situation is the leading edge of a global trend where different jurisdictions impose incompatible requirements. Companies building AI products need geographic flexibility in their model choices—hard-coding a single provider or model creates regulatory risk.

The Bigger Picture

Llama 4 represents Meta's clearest articulation of their AI strategy: commoditize the model layer to prevent competitor lock-in, maintain enough control to protect core revenue streams, and out-invest everyone on pre-training to stay at the frontier.

This strategy only works if Meta continues shipping competitive models. Llama 3 established credibility; Llama 4 maintains it. The test comes with Llama 5 and beyond—can Meta sustain the pace while OpenAI, Anthropic, Google, and an increasingly capable Chinese ecosystem all push forward?

For now, the answer appears to be yes. Meta's compute investments, research talent, and distribution through their product ecosystem create durable advantages. But "for now" is the operative phrase in a field where leadership changes quarterly.

The Bottom Line

Llama 4 delivers meaningful efficiency gains through MoE architecture, genuine multimodal capabilities, and performance that challenges proprietary leaders—at the cost of geographic restrictions that fragment the global AI market.

The practical takeaway: if you’re outside the EU and running AI workloads at scale, Scout deserves immediate evaluation; the infrastructure economics have shifted enough that ignoring it costs money.